AFRL-RI-RS-TR-2010-103

THE SPECTRAL MIXTURE MODELS
A MINIMUM INFORMATION DIVERGENCE APPROACH

April 2010

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■**UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2010-103 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/s/	/s/
STEVEN T. JOHNS, Chief Communications Exploitation Branch	JOSEPH CAMERA, Chief Information & Intelligence Exploitation Division Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*
APRIL 2010

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
October 2008 – November 2009

4. TITLE AND SUBTITLE

THE SPECTRAL MIXTURE MODELS
A MINIMUM INFORMATION DIVERGENCE APPROACH

5a. CONTRACT NUMBER
In House

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62702F

5d. PROJECT NUMBER
459E

5e. TASK NUMBER
H9

5f. WORK UNIT NUMBER
C2

6. AUTHOR(S)

Alfredo Vega Irizarry

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

AFRL/RIEC
525 Brooks Road
Rome, NY 13441-4505

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/RIEC
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TR-2010-103

12. DISTRIBUTION AVAILABILITY STATEMENT
Distribution Approved for Public Release; Distribution Unlimited. PA# 88ABW-2010-2090
Date Cleared: 20-April-2010

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The objective of developing the Spectral Mixture Model Algorithm was to provide some intelligent algorithm that could be utilized for spectral sensing in wideband receivers. The methodology was discussed initially in report AFRL-RI-RS-TR-2008-266. The current report is a refinement of the technique with the objective of presenting the concept to a broader audience. The Spectral Mixture is a generalization of the Expectation Maximization algorithm. The algorithm reduces the information divergence of two distributions by adjusting its parameters. The algorithm can be applied to histogram data or sample points for signal decomposition of multimodal signal in terms of mixture elements. The model was applied to spectral analysis with good success in the one-dimensional case. To achieve better convergence, the algorithm may require the constraint of some of the parameters by imposing boundary conditions or preventing changes. This research explored some potential applications of the algorithm. These include: spectral characterization, speech compression, deconvolution and image processing. The results are summarized in this report.

15. SUBJECT TERMS
Spectral Mixture Model, Expectation Maximization algorithm, spectral characterization, speech compression, deconvolution and image processing.

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
U	U	U	UU	35

19a. NAME OF RESPONSIBLE PERSON
Alfredo Vega Irizarry

19b. TELEPHONE NUMBER (*Include area code)*
N/A

Table of Contents

List of Figures

List of Tables

Abstract

The objective of developing the Spectral Mixture Model Algorithm was to provide some intelligent algorithm that could be utilized for spectral sensing in wideband receivers. The methodology was discussed initially in report AFRL-RI-RS-TR-2008-266. The current report is a refinement of the technique with the objective of presenting the concept to a broader audience.

The Spectral Mixture is a generalization of the Expectation Maximization algorithm. The algorithm reduces the information divergence of two distributions by adjusting its parameters. The algorithm can be applied to histogram data or sample points for signal decomposition of multimodal signal in terms of mixture elements. The model was applied to spectral analysis with good success in the one-dimensional case. To achieve better convergence, the algorithm may require the constraint of some of the parameters by imposing boundary conditions or preventing changes.

This research explored some potential applications of the algorithm. These include: spectral characterization, speech compression, deconvolution and image processing. The results are summarized in this report.

1. Introduction

This report is divided in two parts. The first part is a theoretical discussion of the development of the Spectral Mixture algorithm from the perspective of machine learning and the second part is a discussion of the implementation of the method for some specific applications.

The application of interest is spectral decomposition and characterization, also referred as spectral sensing. Given a spectral estimate, we would like to model an undetermined number of signals in terms of the center frequency, bandwidth and power density. The signals are required to follow some predetermined model in the frequency domain. The method resembles a Gaussian Mixture approach with this important difference: the Gaussian Mixture Model algorithm requires statistical sample points while the Spectral Mixture Model requires histogram data as the input.

The Spectral Mixture algorithm resembles some kind of entropy formulation. By changing the form of the expression, one can prove that the algorithm is an iterative process that reduces the information divergence of two parametric probability densities until the process converges to some local minima. Under certain specific conditions, the algorithm reduces to Expectation Maximization, K-Means, weighted K-Means and Parzen Window depending on the variable is chosen for optimization.

One of the issues of the algorithm is caused by the presence of an offset in the distribution. To overcome this issue, one can assign fixed mixtures to model a signal offset by keeping a fixed width or variance. This issue is cover in section 3.

The algorithm was applied to signal detection and speech compression. The speech compression method has very low complexity and does not require the addition of voiced/unvoiced detection or estimation of the linear prediction coefficients. The price to pay is a reduction of the quality of the speech. The compressed speech sound hoarse, but this is typical of vocoder at a low bitrates.

The report pretends to be an introductory tutorial for those who are interested in this method. This review would serve as baseline for future research in these methods.

2. Classification Algorithms and Metrics

Many classification and machine learning algorithms require a data set of observations, a parametric model and optimization rules. Through an optimization process, we find the parameters that provide a best fit for the model according to the selected rules. The result of this process is called analysis. The analysis is usually the most computationally involved process of classification. Optimization methods require processing many data points, calculating complex derivatives, iterating many times or evaluating complex functions. For such reason, a metric of performance for the analysis process can be described in terms of the number of floating point operations that are needed to find a solution.

Synthesis is the inverse of the analysis process. The synthesis simply requires evaluating the model using certain specified parameters. Measuring the fitness of a model to a data, i.e., calculating the error between the model and the data can be use as a metric of performance. Minimizing the error is a typical criterion for developing optimization rules.

It is often said that the results of the analysis are as good as the model used to fit the data. If the model is not representative of the data, then the results are questionable or perhaps wrong. Various information criteria have been proposed such as the Akaike and Bayesian Information Criterion. Developing a metric that measures the fitness of different models is beyond the scope of our discussion.

2.1.Mixture Models

Suppose that a given data set can be modeled as a linear combination of K density functions. A parameter vector $\vec{\theta} = [\alpha, \mu, \sigma^2, \dots]$ corresponds to the mixture probability α, the mean μ, the variance σ^2 and possibly other high order moments that control the shape of the model. For simplicity, we consider a vector with three parameters $\vec{\theta} = [\alpha, \mu, \sigma^2]$. The parametric model can be express as a summation of mixture terms that follows a given parametric model as

$$p(y|\vec{\theta}) = \sum_{k=1}^{K} \alpha_k p(y_j | \mu_k, \sigma_k^2) \qquad 2.1$$

The maximum likelihood approach provides a criterion for finding the optimal parameter vector. The optimization process finds the parameters that maximize the likelihood function of the model shown in equation 2.2.

$$log\big(p(y|\alpha, \mu, \sigma)\big) = \sum_{j=1}^{J} log\left(\sum_{k=1}^{K} \alpha_k p(y_j | \mu_k, \sigma_k)\right) \qquad 2.2$$

Finding maximum likelihood solutions using equation 2.1 has several problems. The derivatives of the expression with respect to its parameters produce complex terms with no closed form solutions. A second problem occurs when the variance approaches zero; then the normal distribution diverges. This is not a big problem and can be solved by adding a small non-zero element to the variance estimate. In addition, this case is likely to happen when outliers are present. A third problem is called identifiability. This means that K mixtures can be rearranged in $K!$ combinations, so there are $K!$ possible solutions. Once again, the problem is not critical. We cannot rely on the index number that identifies a mixture because it can change its value.

An equivalent and more convenient way to solve the maximum likelihood of a mixture is by using the Expectation-Maximization method shown in equation 2.3:

$$\underset{\theta}{\operatorname{argmax}}\, log\big(p(y|\alpha,\mu,\sigma)\big) \qquad 2.3$$

where the equation is subject to

$$\sum_{k=1}^{K} \alpha_k = 1 \qquad 2.4$$

The differentiation of $log\big(p(Y|\alpha,\mu,\sigma)\big)$ produces a common term in all the expressions for α, μ and σ^2. The term is a posterior distribution:

$$p\big(\alpha_k,\mu_k,\sigma_k^2\big|y\big) = \frac{\alpha_k p\big(y\big|\alpha_k,\mu_k,\sigma_k^2\big)}{\sum_{i=1}^{K} \alpha_i p\big(y\big|\alpha_i,\mu_i,\sigma_i^2\big)} \qquad 2.5$$

The evaluation of equation 2.5 is referred as the expectation step or E-Step.

The solution of equation 2.3 is obtained through a series of iterations that calculate the posterior distribution and maximize the resulting expression in terms of the parameter vector. The M-Step or maximization process consists of the reestimation of the parameter vector. In the case of the well known Gaussian Mixture Model (GMM) algorithm, the model is given by a Gaussian density and the optimization requires the evaluation of equations 2.6-a through 2.6-d.

$$J_k = \sum_{j=1}^{J} p\big(\alpha_k,\mu_k,\sigma_k^2\big|y_j\big) \qquad 2.6\text{-a}$$

$${\alpha_k}^{new} = \frac{J_k}{J} \quad 2.6\text{-b}$$

$$\mu_k{}^{new} = \frac{1}{J_k}\sum_{j=1}^{J} p\left(\alpha_k, \mu_k, \sigma_k^2 \middle| y_j\right) y_j \qquad \text{2.6-c}$$

$$\sigma_k^{2\,new} = \frac{1}{J_k}\sum_{j=1}^{J} p\left(\alpha_k, \mu_k, \sigma_k^2 \middle| y_j\right)(y_j - \mu_k)^2 \qquad \text{2.6-d}$$

In order to go from a maximum likelihood approach to the EM algorithm, we note that the maximum likelihood algorithm provides knowledge of the mixture through the prior distribution $p\left(\alpha_k, \mu_k, \sigma_k^2 \middle| y\right)$. Instead of ignoring this term, we decide to include this knowledge in our advantage by reformulating equation 2.2 such that the summation term does not appear inside the logarithm. This is desirable because it facilitates solving the equations as we will see soon.

The algorithm can be reformulated by creating a dummy variable z. The variable z acts like a tag that provides the index number of a given mixture. The index information is not available directly, so we call this the missing random variable.

$$p\left(y, z = k; \vec{\theta}\right) = \sum_{k=1}^{K} \alpha_i p(y_j \,|z_j = i; \mu_i, \sigma_i^2) \qquad \text{2.7}$$

Some knowledge of the label z is available through $p\left(z = i \middle| y, \alpha_i, \mu_i, \sigma_i^2\right) = p\left(\alpha_i, \mu_i, \sigma_i^2 \middle| y\right)$. This knowledge is supplied into the equation 2.2 by estimating the conditional expectation of the log-likelihood. The new expression is called the auxiliary function $\mathcal{L}$. The optimization rule consists of finding the parameter vector that maximizes the auxiliary function,

$$\mathcal{L}\left(\vec{\theta}, \vec{\theta}_{old}\right) = \sum_{j=1}^{J} p\left(y_j|z_j = k; \vec{\theta}_{old}\right) \sum_{k=1}^{K} log\left(p\left(y_j, z_j = k; \vec{\theta}\right)\right) \qquad \text{2.8}$$

conditioned to $\sum_{k=1}^{K} P\left(z_j = k; \vec{\theta}\right) = 1$.

The E-Step consists of the evaluation of $p\left(z = i \middle| y, \alpha_k, \mu_k, \sigma_k^2\right)$. The M-Step is the maximization of the auxiliary function:

$$\nabla_{\vec{\theta}} \mathcal{L}\left(\vec{\theta}, \vec{\theta}_{old}\right) = 0 \qquad \text{2.9}$$

2.2. Spectral Mixtures 1-D

The Mixture Model approach operates in a set of statistical samples and produces the parameters of a distribution that fit a given model in an optimal sense. A similar approach was developed in project AFRL-RI-RS-TR-2008-266 with the purpose of analyzing histograms instead of statistical samples. The objective was to create a method for spectral sensing, i.e., a method for characterizing the frequency spectrum of telecommunication signals.

The spectral model can be derived from the quantization of the feature space. Given a data set, its quantized samples generate a histogram. The histogram is simply a pair of cells and sample counts. The cell indexes will substitute the data set in the EM method. The sample count will be reflected as exponents in the likelihood function.

First, we define continuous variable $\{\, y \mid y \in \mathbb{R} \,\}$, a discrete variable an arbitrary $\{\, f \mid f \in \mathbb{Z} \,\}$ and a quantization function $Q(y)$ that maps y into a discrete variable f:

$$Q\colon \{f = Q(y) \mid y \in \mathbb{R}, f \in \mathbb{Z}\} \qquad 2.10$$

Let's assume that finite set of statistical samples $\{y_j\}_{j=1}^{J}$ is sorted in ascending order such that $y_j < y_k \; for \; j < k$. The quantized process forms a new set $\{Q(y_j)\}_{j=1}^{J}$. We are interested in counting the number of repeated values in the new set. We define S as the count of samples for a quantization index f as,

$$S(f) = \; Count\left\{f\colon f = \{Q(y_j)\}_{j=1}^{J}\right\} \qquad 2.11$$

The goal is to investigate the form of the auxiliary function (equation 2.8) under quantization of the sample space. Quantization always produces lost of information. In this case, specific information of the missing variable z is expected to be lost for individual samples. Under quantization, all samples falling under the same cell i must share the same index number. We designate a new variable $\bar{z}$ to represent a tag for group of samples within the same histogram bin. We also impose the probability function to be identical for all samples inside a cell.

$$P(\bar{z} = i) \equiv P(z_j = i) \qquad 2.12$$

The likelihood $p\left(f \middle| \bar{z} = k; \vec{\theta}\right)$ will be defined as an average of the probability models. This is done by adding all likelihoods with a common quantization index f and dividing them by the number of samples in the histogram bin.

$$p\left(f \middle| \bar{z} = k; \vec{\theta}\right) \equiv \frac{1}{S(f)} \sum_{j=1}^{J} p\left(f = Q\{y_j\} \middle| z_j = k; \vec{\theta}\right) \quad 2.13$$

The posterior probability $p(y_j|z_j = k; \vec{\theta}_{old})$ remains invariant under the transformation of equation 2.13.

$$p(\bar{z} = k|f; \vec{\theta}_{old}) = \frac{\alpha_k p(f|\bar{z} = k; \vec{\theta}_{old})}{\sum_{k=1}^{K} \alpha_k p(f|\bar{z} = k; \vec{\theta}_{old})} \quad 2.14$$

Finally, let's rewrite the auxiliary function under quantization. Each repeated samples modify the likelihood expression by rising each term to the $S(f_m)$ power:

$$\mathcal{L}(\vec{\theta}, \vec{\theta}_{old}) = \sum_{m=1}^{M} p(\bar{z}_m = k|f_m; \vec{\theta}_{old}) \sum_{k=1}^{K} log\left(p(f_m, \bar{z}_m = k; \vec{\theta})^{S(f_m)}\right) \quad 2.15$$

subject to the constraint $\sum_{k=1}^{K} P(\bar{z}_j = k; \vec{\theta}) = 1$.

It was found convenient to rename the parameter vector using variables associated with the frequency spectrum, which in a sense; it is nothing less than a histogram. The terms mean, variance and mixture probability will be replaced by center frequency f_c, statistical bandwidth b, and power composition ρ respectively. The parameter vector is rewritten as:

$$\vec{\theta} = [\rho, f_c, b] \quad 2.16$$

2.2.1. Maximization of the Expectation

This section covers the discussion of the maximization process for the one and two-dimensional case. The one-dimensional case initial motivation was to analyze the frequency spectrum to characterize signals above the noise floor. The motivation for the two-dimensional case analysis was to analyze signals in the time-frequency domain. The analysis resulted in a rudimentary and perhaps inefficient image detector. The results of the findings will be summarized in the present section.

Let's now define a family of Gaussian-like density functions as,

$$p(f|z; \vec{\theta}) = \frac{N}{b\,\Gamma\left(\frac{1}{N}\right)} exp\left\{-\left(\frac{f - f_c}{b/2}\right)^N\right\} \quad 2.17$$

For $N = 2$, the distribution becomes Gaussian with mean f_c and variance $b^2/8$. The function $\Gamma(x)$ is the Gamma function. The model has been plotted for several values of N in Figure 1.

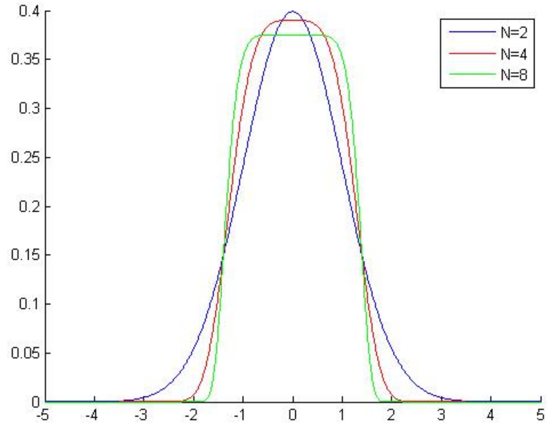

Figure 1 Parametric model of equation 2.17 as a function of N

Maximizing the expectation results in equations 2.18-a through 2.18-d.

$$\frac{\partial}{\partial f_{c,k}}\mathcal{L}\big(\vec{\theta},\vec{\theta}_{old}\big)=\sum_{m=1}^{M}p\big(\bar{z}_m=k|f_m;\vec{\theta}_{old}\big)S(f_m)\ \frac{N}{(f_m-f_c)}\left(\frac{f_m-f_{c,k}}{b_k/2}\right)^{N}=0 \qquad 2.18\text{-a}$$

$$\frac{\partial}{\partial b_k}\mathcal{L}\big(\vec{\theta},\vec{\theta}_{old}\big)=\sum_{m=1}^{M}p\big(\bar{z}_m=k|f_m;\vec{\theta}_{old}\big)S(f_m)\left(-\frac{1}{b_k}-\frac{N}{(b_k/2)}\left(\frac{f_m-f_{c,k}}{b_k/2}\right)^{N}\right)=0 \qquad 2.18\text{-b}$$

$$\frac{\partial}{\partial \rho_k}\mathcal{L}\big(\vec{\theta},\vec{\theta}_{old}\big)=0\ \rightarrow P(\bar{z}_m=k)=\frac{p\big(\bar{z}_m=k|f_m;\vec{\theta}_{old}\big)\,S(f_m)}{\sum_{i=1}^{K}\sum_{n=1}^{M}p\big(\bar{z}_n=i|f_n;\vec{\theta}_{old}\big)\,S(f_n)} \qquad 2.18\text{-c}$$

The optimal parameter vector $\vec{\theta}_{opt}=[\rho_{opt},f_{c_{opt}},b_{opt}]$ is obtained by implementing the following steps:

Algorithm:

1. E-Step: Evaluate $p\big(\bar{z}_m=k|f_m;\vec{\theta}_{old}\big)$
2. M-Step: Evaluate $\vec{\theta}=[\rho,f_c,b]$ using equations 2.18-a to 2.18-c
3. Verify convergence by comparing $\vec{\theta}$ and $\vec{\theta}_{old}$

2.2.2. Spectral Mixtures 2-D

An interesting development of the spectral mixtures is when extending the concept for the two-dimensional case. We are going to show how quantities like as center of mass, standard deviation and regression line arises from the maximization of our Gaussian-like model.

The parametric model used is a Clipped Gaussian distribution in two dimensions. The distribution is controlled by a parameter N that modifies the decaying slope of the density. We also introduce a set of affine transformations modified by the following parameters: translation, scale, shear and rotation. In our case, shear is constrained by scale because the intention of this particular model is to characterize modulated signals in the joint time-frequency domain.

Our 2-D model defines an affine transformation Q, a translation vector $\vec{\mu}$ and a coordinate vector $\vec{x}$. The affine transform Q includes a shear matrix $[Sh]$, a scale matrix $[Sc]$ and a rotation matrix $[R]$.

$$Q(\vec{x}-\vec{\mu}) = [Sh]^{-1}[Sc]^{-1}[R]^{-1}(\vec{x}-\vec{\mu}) \quad 2.19$$

The parameters of the matrices are the angle of rotation ϕ, the scale factor s_x in the x coordinate and the scale factor s_y in the y coordinate. The parameter $\vec{\mu}$ is a translation vector. The specific form of the matrices is shown in equations 2.20 through 2.22.

$$R(\phi) = \begin{bmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{bmatrix} \quad 2.20$$

$$Sc = \begin{bmatrix} \mathrm{s_x} & 0 \\ 0 & \mathrm{s_y} \end{bmatrix} \quad 2.21$$

The model provides sufficient parameters for characterizing telecommunication signals in the time-frequency domain. The shear models a parallelogram shape which is characteristic of chirp signals in the joint time-frequency domain.

$$Sh = \begin{bmatrix} 1 & 0 \\ -\frac{\mathrm{s_y}}{\mathrm{s_x}}\tan(\phi) & 1 \end{bmatrix} \quad 2.22$$

A new vector is constructed by raising each coordinate to the N power. Equation 2.23 is a non-linear transformation.

$$q = \{\, Q(\vec{x}-\vec{\mu})\,\}^{N} \quad 2.23$$

The resulting exponential function resembles a rectangular surface for $N = 4$ as shown in Figure 2.

$$p(\vec{x}|z;\vec{\theta}) = \frac{1}{V(\vec{\theta})} exp\{-q^T q\} \quad 2.24$$

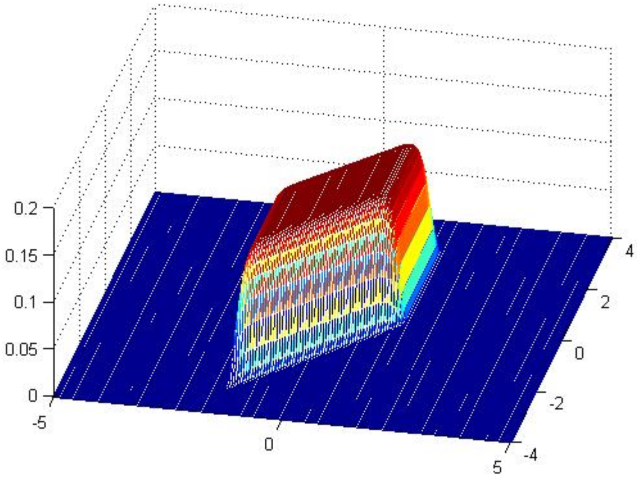

Figure 2 Shape of the parametric model in 2 dimensions of equation 2.20

The parameter vector is defined in terms of these geometrical parameters as,

$$\vec{\theta} = [\rho, \mu_x, \mu_y, s_x, s_y, \phi] \quad 2.25$$

where ρ is the ratio of the volume of one mixture over the entire volume.

Normalization requires integrating the whole surface. For $N = 4$, the volume was approximated using equation 2.26. The volume of the distribution is a function of to the scales and the order of the distribution N. For our purposes, N is constant so the volume can be expressed as a function of the scaling parameters.

$$V(\vec{\theta}) = \iint exp\{-q^T q\}\, dx\, dy \ \approx c \cdot sx \cdot sy \quad 2.26$$

The auxiliary function $\mathcal{L}$ keeps the same appearance. The variable f_m is replaced with vector $\vec{x}$:

$$\mathcal{L}(\vec{\theta}, \vec{\theta}_{old}) = \sum_{m=0}^{M} p(\bar{z}_m = k|\vec{x}_m; \vec{\theta}_{old}) \sum_{k=1}^{K} log\left(p(\vec{x}_m, \bar{z}_m = k; \vec{\theta})^{S(\vec{x}_m)}\right) \qquad 2.27$$

subject to the constraint $\sum_{k=1}^{K} P(\bar{z}_m = k; \vec{\theta}) = 1$.

Maximizing the expectation requires finding the roots of the derivatives of the auxiliary function. The derivatives can be found by using *Mathematica* or other symbolic math application.

For the translation μ_x, the derivatives are given by equation 2.28.

$$\begin{aligned}\frac{\partial}{\partial\mu_x}\mathcal{L}(\vec{\theta},\vec{\theta}_{old}) = &\sum_x\sum_y p\left(\bar{z}_m = k|x_m, y_m; \vec{\theta}_{old}\right)\cdot S(x_m, y_m)\\ &\cdot\left\{ -4^N\left(\left(\frac{(x-\mu_x)\text{sec}\,(\phi)}{s_x}\right)^N + \left(\frac{(x-\mu_x)\cos(\phi) + \left(y-\mu_y\right)\cos\,(\phi)}{s_y}\right)^{2N}\right)\right.\\ &\left.-\ln\left(\mathrm{k}\cdot \mathrm{s_x}\cdot \mathrm{s_y}\right)\right\} = 0\end{aligned} \qquad 2.28$$

Some simplification can be accomplished by using $N = 1$ as an approximation.

$$\mu_x = \frac{\sum_x\sum_y p\left(\bar{z}_m = k|x_m, y_m; \vec{\theta}_{old}\right)\cdot S(x_m, y_m)\cdot x}{\sum_x\sum_y p\left(\bar{z}_m = k|x_m, y_m; \vec{\theta}_{old}\right)\cdot S(x_m, y_m)} \qquad 2.29$$

This is an expression equivalent to the center of mass in the x coordinate. A similar expression was found for the y coordinate.

$$\mu_y = \frac{\sum_x\sum_y p\left(\bar{z}_m = k|x_m, y_m; \vec{\theta}_{old}\right)\cdot S(x_m, y_m)\cdot y}{\sum_x\sum_y p\left(\bar{z}_m = k|x_m, y_m; \vec{\theta}_{old}\right)\cdot S(x_m, y_m)} \qquad 2.30$$

For the scaling s_x, the maximization of the auxiliary function produces a closed form solution.

$$s_x = \left(\frac{\sum_x\sum_y p\left(\bar{z}_m = k|x_m, y_m; \vec{\theta}_{old}\right)\cdot S(x_m, y_m)\cdot \{2^{1+2N}N(\sec(\phi))^{2N}(x-\mu_x)^{2N}\}}{\sum_x\sum_y p\left(\bar{z}_m = k|x_m, y_m; \vec{\theta}_{old}\right)\cdot S(x_m, y_m)}\right)^{1/(2N)} \qquad 2.31$$

In a similar manner, the scaling s_y is given by,

$$\begin{aligned}&s_y\\ &= \left(\frac{\sum_x\sum_y p\left(\bar{z}_m = k|x_m, y_m; \vec{\theta}_{old}\right)\cdot S(x_m, y_m)\cdot \left\{2^{1+2N}N\left((x-\mu_x)\sin(\phi) - \left(y-\mu_y\right)\cos(\phi)\right)\right\}}{\sum_x\sum_y p\left(\bar{z}_m = k|x_m, y_m; \vec{\theta}_{old}\right)\cdot S(x_m, y_m)}\right)^{1/(2N)}\end{aligned}$$

2.32

These expressions reduce to the standard deviation when the order of the distribution is one. In the y coordinate, the standard deviation is calculated along the regression line of the data given by $(x - \mu_x) \sin(\phi) - (y - \mu_y) \cos(\phi) = 0$. One can argue that the standard deviation provides a rough approximation of the scales s_x and s_y during our optimization process. This approximation was verified during simulations.

For the angle of rotation ϕ, the maximization of the auxiliary function results in equation 2.33.

$$\sum_x \sum_y p(\bar{z}_m = k | x_m, y_m; \vec{\theta}_{old}) \cdot S(x_m, y_m) \cdot \left\{-2N\left(\left(\vec{w}_{||} \cdot (\vec{p} - \vec{p}_0)\right)^{2N-1} (\vec{w}_\perp \cdot (\vec{p} - \vec{p}_0))/s_y^{2N}\right) + \left((x - \mu_x)\sec(\phi)\right)^{2N} \tan(\phi)\right\} \qquad 2.33$$

The vectors $\vec{w}_{||}$, $\vec{w}_\perp$, $\vec{p}$ and $\vec{p}_0$ are vectors the principal components of the data as illustrated in Figure 3. The data is distributed inside a parallelogram that is at a distance $\vec{p} - \vec{p}_0$ from the origin. The slope of the principal component equals $\tan(\phi)$ and runs parallel to the regression line $\vec{w}_{||}$. The dot product of the distance vector and the principal component $\vec{w}_\perp$ defines the regression line.

$$\vec{w}_{||} \cdot (\vec{p} - \vec{p}_0) = -\sin(\phi)(x - \mu_x) + \cos(\phi)(y - \mu_y) = 0 \qquad 2.34$$

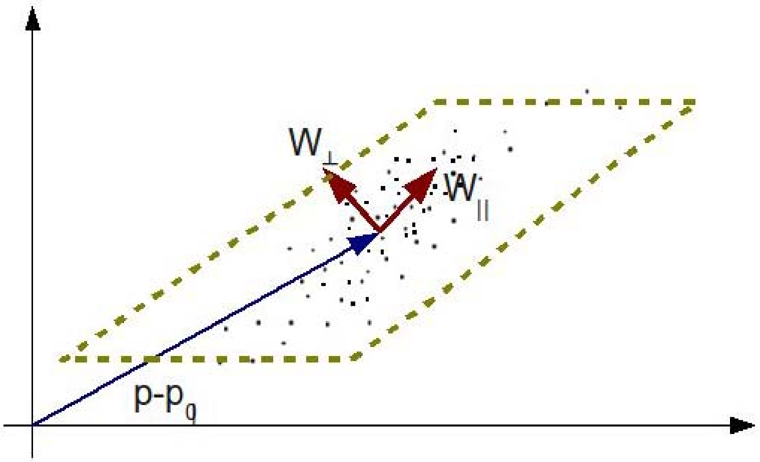

Figure 3 Distribution of data points of the selected model: Two principal component $w_\perp$ and $w_{||}$ from distribution of equation 2-20.

The regression line approximation is verified when $N = 1$. The slope of the regression line becomes:

$$\tan\,(\phi) = \frac{\sum_x \sum_y p\big(\bar{z}_m = k | x_m, y_m; \vec{\theta}_{old}\big) \cdot S(x_m, y_m) \cdot (x - \mu_x) \cdot \big(y - \mu_y\big)}{\sum_x \sum_y p\big(\bar{z}_m = k | x_m, y_m; \vec{\theta}_{old}\big) \cdot S(x_m, y_m) \cdot \big(y - \mu_y\big)} \quad 2.35$$

The optimal parameter vector $\vec{\theta}_{opt}$ is obtained by implementing the following steps.

Algorithm 2-D:

1. E-Step: Evaluate $p\big(\bar{z}_m = k | x_m, y_m; \vec{\theta}_{old}\big)$
2. M-Step: Evaluate $\vec{\theta} = [\rho, \mu_x, \mu_y, s_x, s_y, \phi]$ using equations 2.25 to 2.31.
3. Verify convergence by comparing $\vec{\theta}$ and $\vec{\theta}_{old}$

2.2.3. Relationship between Spectral Mixtures and Parzen Windows

Parzen Windows is a non-parametric method used for the estimation of probability density function. In machine learning, a non-parametric method implies that there is no control over the model. The most common way for adjusting the parameters of a Parzen Window is by trial and error. The best estimate comes from reducing the error between the data histogram and the model. This can be done by visual inspection. Other approaches use neural networks to accomplish this task.

The method models discrete histogram bins by means of kernel functions. In the simplest case, it defines a rectangular window of unit area and width h. The density $p(x_i)$ is estimated using the window function $w(x)$ shown in Figure 4 as blue rectangles, the number of data points x_i that falls under the window and the area (or volume) of the window.

$$p(x_i) = \frac{samples\ inside\ the\ bin\ i}{total\ samples \cdot area} \quad 2.36$$

$$p(x_i) = \frac{\sum_{i=1}^{K} w\left(\frac{x_m - x_i}{h_M}\right)}{K \cdot h} \quad 2.37$$

$$w(x) = \begin{cases} 1 & for\ -\frac{1}{2} \leq x < \frac{1}{2} \\ 0 & otherwise \end{cases} \quad 2.38$$

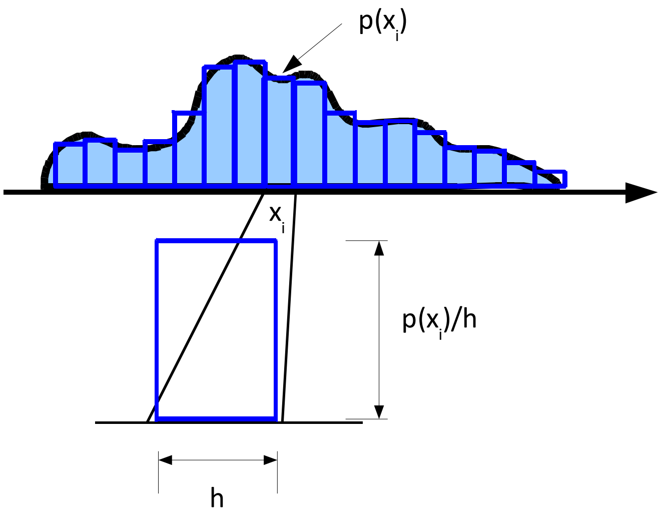

Figure 4 Concept of Non-Parametric Parzen Windows

The window $w(x)$ is known as the kernel. It does not need to be a rectangular window. Parzen proved that it can take the form of equation 2.39 or other shapes as long as the following conditions are met:

$$\int_{-\infty}^{\infty} |\, w(x) \,|\, dx \;\geq 1, \quad 2.39$$

$$|\, w(x) \,| \;\leq\; \infty, \quad 2.40$$

$$\lim_{x \to \infty} |\, x \cdot w(x) \,| = 0. \quad 2.41$$

It is possible for this type of kernel to converge to the original density function when $h \;\rightarrow 0$ for a large number of windows is a normalized density function:

$$\int_{-\infty}^{\infty} w(x)\, dx = 1 \quad 2.42$$

The Spectral Mixture algorithm is also based in quantization of the sample space. This similarity between Parzen Windows and Spectral Mixtures make one wonder if there is a relationship between both algorithms. In the spectral mixture case, the approximate number of samples in bin x_m is given by:

$$S(x_m) = \sum_{i=1}^{M} w\left(\frac{x_m - x_i}{h_M}\right) \quad 2.43$$

These samples modify the likelihood in the form of exponents as shown previously.

In the Parzen Window method, the bandwidth b_i and the center frequency x_i are at fixed. The minimization process is done with respect to the parameter $\rho_k^{new} = P(\bar{z}_m = k)$. We also impose that the number of samples and mixtures are the same ($M = K$). These assumptions are incorporated in equations 2.44 and 2.46 as:

$$\mathcal{L}(\rho_{new}, \rho_{old}) = \sum_{m=1}^{M} p(\bar{z}_m = k|x_m; \rho_{old}) \sum_{k=1}^{K} log\left(p(x_m, \bar{z}_m = k; \rho_{new})^{S(x_m)}\right) \qquad 2.44$$

subject to the constraint $\sum_{k=1}^{K} \rho_k^{new} = 1$ and posterior:

$$p\left(\bar{z}_m = k|x_m; \rho_k^{old}\right) = \frac{p\left(x_m\middle|\bar{z}_m = k; \rho_k^{old}\right)\rho_k^{old}}{\sum_{l=1}^{K} p\left(x_m\middle|\bar{z}_m = l; \rho_k^{old}\right)\rho_l^{old}} \qquad 2.45$$

The optimal solution for ρ is given by equation 2.18-c:

$$\rho_k^{new} = \frac{\sum_{m=1}^{M} p\left(\bar{z}_m = k|x_m; \rho_k^{old}\right) S(x_m)}{\sum_{i=1}^{K}\sum_{n=1}^{M} p\left(\bar{z}_n = i|x_n; \rho_k^{old}\right) S(x_n)} \qquad 2.46$$

Substituting equations 2.45 in 2.46 result in 2.47.

$$\rho_k^{new} = \frac{1}{\sum_{n=1}^{M} S(x_n)} \sum_{m=1}^{M} \frac{p\left(x_m\middle|\bar{z}_m = k; \rho_k^{old}\right)\rho_k^{old}}{\sum_{l=1}^{K} p\left(x_m\middle|\bar{z}_m = l; \rho_k^{old}\right)\rho_l^{old}} S(x_m) \qquad 2.47$$

The likelihood has the same form as the normalized window function in 2.37 as:

$$\rho_k^{new} = \frac{1}{\sum_{n=1}^{M} S(x_n)} \sum_{m=1}^{M} \frac{S(x_m) w\left(\frac{x_m - x_k}{h_M}\right)\rho_k^{old}}{\sum_{l=1}^{K} w\left(\frac{x_m - x_l}{h_M}\right)\rho_l^{old}} \qquad 2.48$$

The convergence of the parameter $\rho_k^{new} \to \rho_k^*$ suggests that $1/M$ is a possible solution. This solution is in agreement with the Parzen Window method.

$$\rho_k^* = \frac{1}{\sum_{n=1}^{M} S(x_n)} \sum_{m=1}^{M} \frac{S(x_m) w\left(\frac{x_m - x_k}{h_M}\right)\rho_k^*}{\sum_{l=1}^{K} w\left(\frac{x_m - x_l}{h_M}\right)\rho_l^*} \qquad 2.49$$

The sum over all ρ_k^* must be equal to one. This was our original constraint in equation 2.39, which is also true for the equation 2.50. This shows that the spectral mixture algorithm reduces to Parzen Window when the parameters are fixed and the parameter ρ_k converges to the specified value.

$$\rho_k^* = \frac{1}{M} = \frac{1}{\sum_{n=1}^{M} S(x_n)} \sum_{m=1}^{M} \frac{S(x_m) w\left(\frac{x_m - x_k}{h_M}\right)}{\sum_{l=1}^{K} w\left(\frac{x_m - x_l}{h_M}\right)} \qquad 2.50$$

The Parzen Window may require finding optimal estimates of the bandwidth h_M. The Spectral Mixture approach already provides a way to readjust the bandwidth and distances between mixtures.

2.2.4. K-Means and Spectral Mixtures

The K-Means algorithm is a special case of the Expectation Maximization method for a mixture of Gaussian densities, also known as Gaussian Mixture Models (GMM). The difference between the K-Means and the GMM is that K-means makes a deterministic assignment of samples to a cluster (or hard decision) while the GMM makes a probabilistic assignment or soft decision using a Gaussian distribution instead of a rectangular window function.

Going from a soft decision to a hard decision is equivalent to fixing the variance and allowing finding the limit as it goes to zero. The maximization of the auxiliary function $2\sigma^2 \cdot \mathcal{L}(\mu, \mu_{old})$ is done by with respect to the centroids μ_k.

$$\sigma^2 \cdot \mathcal{L}(\mu, \mu_{old}) = \sum_{m=1}^{M} p(\bar{z} = k|y; \mu_{old}) \sum_{k=1}^{K} \frac{(y - \mu_k)^2}{2} + \sigma^2 \cdot log\left(\frac{\rho_k}{\sqrt{2\pi}}\right) - \sigma^2 \cdot log(\sigma) \qquad 2.51$$

$$\lim_{\sigma^2 \to 0} \left\{ \frac{\partial}{\partial \mu_k} 2\sigma^2 \cdot \mathcal{L}(\vec{\theta}, \vec{\theta}_{old}) \right\} \qquad 2.52$$

As the variance goes to zero, the conditional probability $p(\bar{z} = k|f; \vec{\theta}_{old})$ converges to the unity only at the data point that is closest to the centroid μ_k.

$$\lim_{\sigma_k^2 \to 0} p(\bar{z} = k|y; \mu_{old}) = \gamma(\bar{z} = k|y) = \begin{cases} 0 & for\ y - \mu_k \neq \min(y - \mu_k) \\ 1 & for\ y - \mu_k = \min(y - \mu_k) \end{cases} \qquad 2.53$$

The auxiliary function becomes:

$$2\sigma^2 \cdot \mathcal{L}(\mu, \mu_{old}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{1}{2} \gamma(\bar{z} = k|y)(y - \mu_k)^2 + constants\ terms \qquad 2.54$$

The extending K-Means to Spectral Mixtures is very simple. The distance between the samples f and the bins $f_{c,k}$ are weighted by the histogram count $S(f)$. The resulting algorithm finds K centers of mass of a histogram.

$$2\sigma^2 \cdot \mathcal{L}\big(f_c, f_{c,old}\big) = \sum_{m=1}^{M}\sum_{k=1}^{K}\frac{1}{2}\gamma(\bar{z} = k|f)S(f)(f - f_{c,k})^2 + constants\ terms \qquad 2.55$$

Table 1 Spectral Mixture Cases

Spectral Mixture Case:	*Algorithm:*
Case $(f) = 1$:	Expectation Maximization
Case $S(f) = 1, \sigma \to 0$	K-Means
Case $S(f) = const, f = const,\ \sigma = const.$	Parzen Window

2.3.Spectral Mixtures and Kullback-Leibler Divergence

As we have seen, the Spectral Mixture approach can be considered as a general case of EM, K-Means and Parzen Window. In addition, we can observe that the auxiliary function looks somewhat similar to some sort of entropy.

The Spectral Mixture algorithm will be modified without altering the final result. Instead of using an arbitrary histogram, we decide to normalize $S(\vec{x}_m)$ and expressed it as a probability density function $\bar{p}(\vec{x}_m) = S(\vec{x}_m)/\int S(\vec{x})dV$. The normalization constant $\int S(\vec{x})dV$ and the constant expression $\sum_{m=0}^{M} p\big(\bar{z}_m = k|\vec{x}_m;\vec{\theta}_{old}\big)\bar{p}(\vec{x}_m)\sum_{k=1}^{K} log\left(p\big(\bar{z}_m = k|\vec{x}_m;\vec{\theta}_{old}\big)\bar{p}(\vec{x}_m)\right)$ do not affect the maximization process with respect to $\vec{\theta}$.

$$\begin{aligned}\mathcal{L}\big(\vec{\theta},\vec{\theta}_{old}\big) = &\sum_{m=0}^{M} p\big(\bar{z}_m = k|\vec{x}_m;\vec{\theta}_{old}\big)\bar{p}(\vec{x}_m)\sum_{k=1}^{K} log\left(p\big(\vec{x}_m|\bar{z}_m = k;\vec{\theta}\big)p(\bar{z}_m = k)\right) \\ &+ \sum_{m=0}^{M} p\big(\bar{z}_m = k|\vec{x}_m;\vec{\theta}_{old}\big)\bar{p}(\vec{x}_m)\sum_{k=1}^{K} log\left(p\big(\bar{z}_m = k|\vec{x}_m;\vec{\theta}_{old}\big)\bar{p}(\vec{x}_m)\right)\end{aligned} \qquad 2.56$$

The resulting expression is the negative of the Kullback-Leibler divergence between distributions $q\big(\vec{x},z;\vec{\theta}_{old}\big) = \sum p\big(\vec{x}|\bar{z};\vec{\theta}_{old}\big)\bar{p}(\vec{x})$ and $p\big(\vec{x},z;\vec{\theta}\big) = \sum p\big(\vec{x}|\bar{z};\vec{\theta}\big)p(\bar{z})$.

$$D_{KL}\left(q\big(\vec{x},z;\vec{\theta}_{old}\big)\,||\,p\big(\vec{x},z;\vec{\theta}\big)\right) = -\mathcal{L}\big(\vec{\theta},\vec{\theta}_{old}\big) \qquad 2.57$$

The Spectral Mixture algorithm minimizes the divergence of the distributions corresponding to a selected model and the data distribution. Upon successive iterations, the probability of the data predicted by the model approximates the true distribution.

$$\vec{\theta}_{new} = \ arg \min_{\vec{\theta}} D_{KL}\left(q\big(\vec{x}, z; \vec{\theta}_{old}\big) \,|\,|\, p\big(\vec{x}, z; \vec{\theta}\big)\right) \quad 2.58$$

The minimization process requires the correct model to achieve the following convergence:

$$p(\vec{x}_m) = \sum_{k=1}^{K} p\big(\vec{x}_m | \bar{z}_m = k; \vec{\theta}\big) p(\bar{z}_m = k) \ \rightarrow \ \bar{p}(\vec{x}_m) \quad 2.59$$

Other divergence formulations can be used instead of equation 2.58. The minimization of the parametric Renyi alpha divergence [9] (equation 2.60) will be based on the log of the ratio between the model and the data. This is a maximum likelihood formulation analogous to the equation 2.2 that originated all this theory.

$$D_{\infty}\left(q\big(\vec{x}, z; \vec{\theta}_{old}\big) \,|\,|\, p\big(\vec{x}, z; \vec{\theta}\big)\right) = \log\left(\frac{p\big(\vec{x}, z; \vec{\theta}\big)}{q\big(\vec{x}, z; \vec{\theta}_{old}\big)}\right) \geq 0 \quad 2.60$$

3. Implementation of the Spectral Mixture Algorithm

3.1.Spectral Sensing

The Spectral Mixture model algorithm was implemented in Matlab using 1-D Gaussian densities and provided as an example. Some rules were implemented to control the convergence of the mixtures in the M-Step. For instance, if the variance is less than a minimum value, then reset the variance to a predetermined value.

The first case is an example of a pure Gaussian Mixture process with $N = 2$ in equation 2.17. Four Gaussian mixtures were generated with noise. (See Figure 5) A total of ten mixture elements were added and the algorithm iterated 20 times.

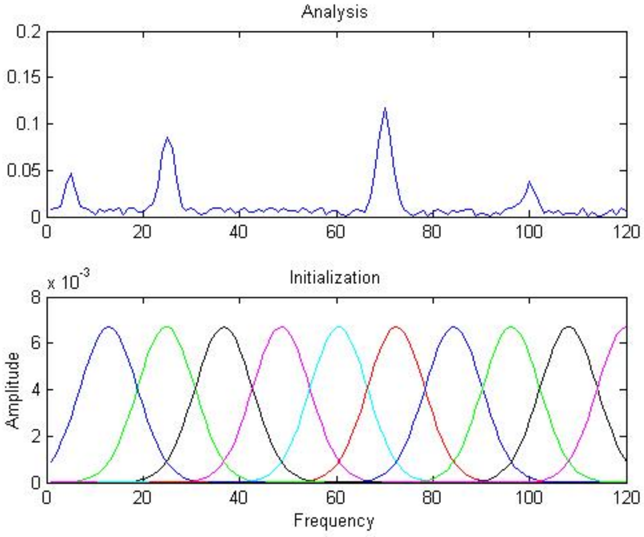

Figure 5 Analysis of a histogram of Gaussian mixtures

Convergence to the desired values was achieved in the first 5 iterations approximately as shown in Figure 6.

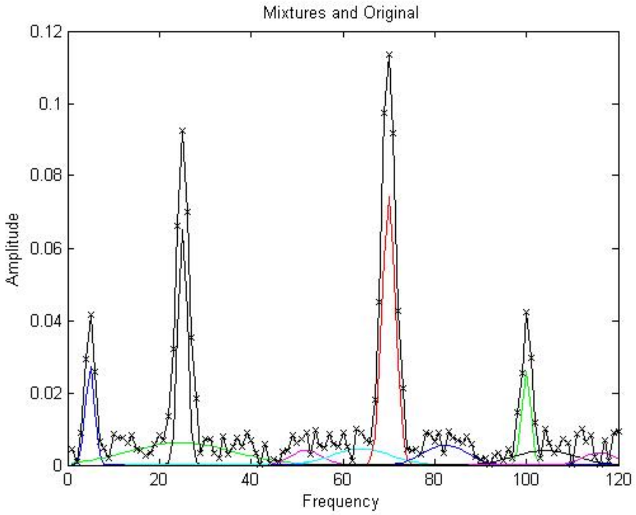

Figure 6 Convergence of the Gaussian mixtures

The overall contribution of all the mixtures has an average squared error of 0.009. The synthesis shown in Figure 7 resembles the original distribution. One of the explored ideas was the design of a simple vocoder that encodes the values of the mean and variances and then synthesizes the speech. This will be discussed in section 3.2.

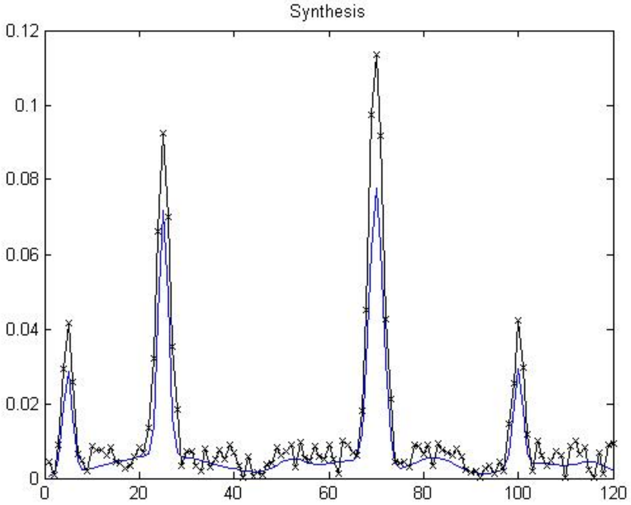

Figure 7 Synthesis of the Spectral mixtures

The spectral mixtures work for clipped Gaussian distributions generated from equation 2.17. In this case, the feasibility of using the algorithm for spectral sensing applications was explored. For these experiments, a polyphase filter was used to provide a spectral average of generic MPSK and FPSK signals. Figure 8 shows two MPSK signals under SNR below 15 dB.

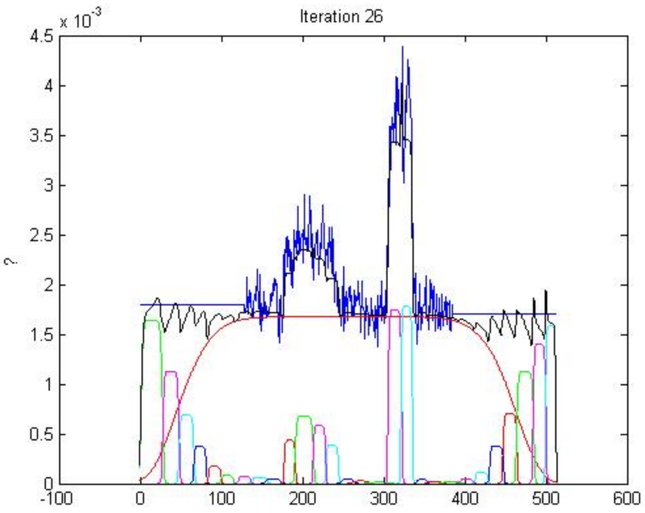

Figure 8 Synthesis of the frequency spectrum

The algorithm achieves convergence when one of the mixtures is constrained from moving and expanding: fixed center frequency and fixed bandwidth. The trick forces the mixture to converge to the noise floor of the spectrum while the other mixtures converge to the signals above the noise floor.

An attempt to implement a two-dimensional Spectral Mixture revealed that the process is slow, computationally intensive and does not converge so easily to the desired solutions.

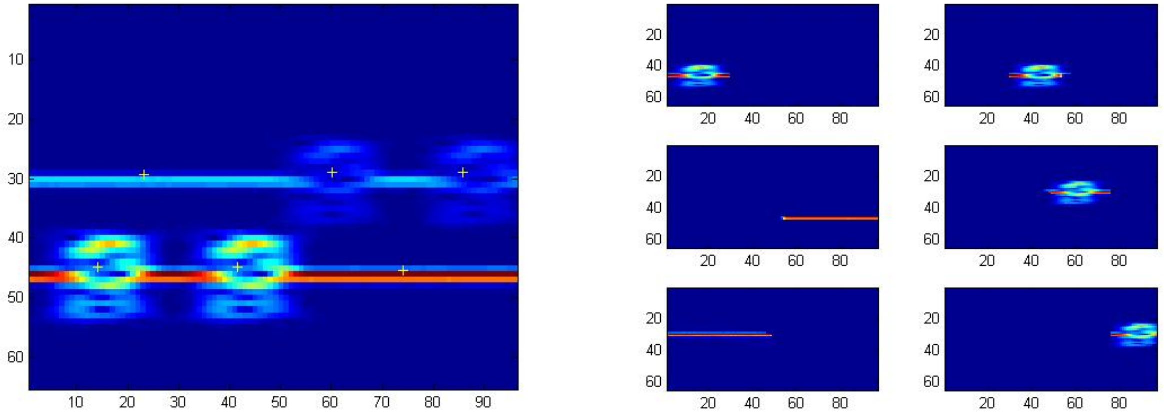

Figure 9 Convergence of a 2-dimension Spectral Mixture, 45dB

Figure 9 shows two BPSK signals switching from on and off states in a short time period. The vertical axis represents frequency and the horizontal one represents time. The SNR of the signals is above 45 dB. The figure to the left shows the centroids after convergence. The right-side figure shows the product of the posterior distribution $p(\bar{z}_m = k|\vec{x}_m; \vec{\theta}_{old})$ with the spectral distribution $S(f)$. The posterior distributions act as a window that block the undesired signals and only allow the portion of the distribution $S(f)$ that contributes to a mixture.

Under more severe SNR and without the usage of a fixed mixture that converges to the noise floor, the method breaks. The centroids are attracted by the areas of high density. In some cases, a mixture can converge to large of portions of the noise floor as shown in Figure 10. The addition of the fixed mixture shows very little improvement.

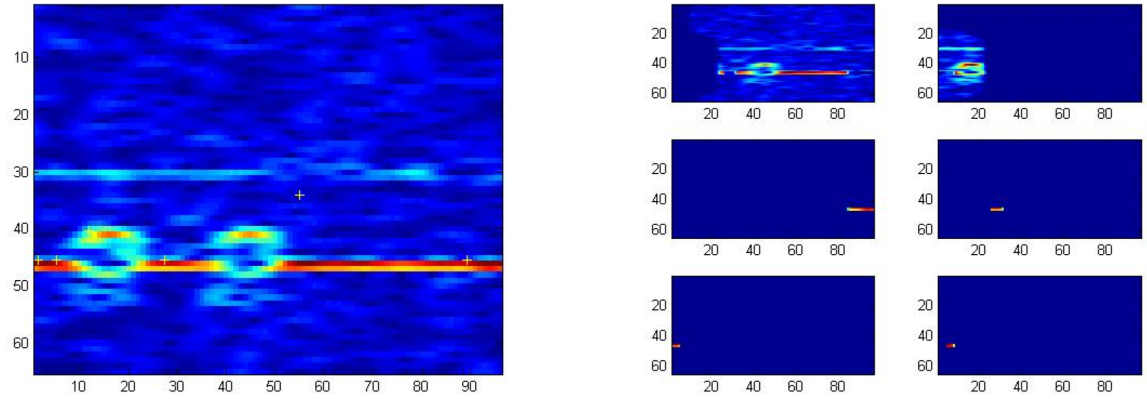

Figure 10 Convergence of a 2-dimension Spectral Mixture, 6dB

3.2.Simple Vocoder

Vocoders are a set of coders specifically designed for voice. Basically there are three types of speech coding: waveform coding (such as ADPCM), source or parametric coders (vocoders) and hybrid coders which are a combinations of the first two. One important difference among these approaches is the quality and achievable bitrate. The hybrid coders have the best quality and performance, but also they are the ones with the highest complexity.

Vocoders rely on speech synthesis and psychoacoustic models for speech synthesis. The vocoders rely on analysis and synthesis that produce speech that is perceptually acceptable. The synthetic speech does not have to necessary match the original signal. The synthetic speech is usually as good as the model. The model usually separates a speech signal into voiced or periodic speech and unvoiced speech. The signals are also separated in the excitation and the vocal track response. The problem of characterizing the vocal track is related to linear prediction. The linear prediction problem is equivalent to a system identification problem and it can be proved that the optimal linear prediction coefficients that minimize the error are exactly the coefficients of autoregressive model that generates speech. [2]

The Multiband Excitation Vocoder is an "analysis by synthesis" method which means that the synthetic speech is compared with the original to get an error signal. [8] The method minimizes the error by adjusting the spectral envelope. The model also ignores the phases of the spectral response as shown in equation 3-1. The MBE defines a criterion error to be minimized. The error is just the difference between the spectral power of the original $S(\omega)$ and the synthetic signal $\hat{S}_w(\omega)$ which is the product of the spectral response given the vocal track $H_w(\omega)$ and the excitation $E_w(\omega)$.

$$Error = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S(\omega)| - |H_w(\omega)||E_w(\omega)||^2 d\omega \quad 3.1$$

The Spectral Mixture provides a way to analyze and synthesize the spectral response $|S(\omega)|$ of the speech. By using a nearly perfect reconstruction filter bank and Spectral Mixtures we can encode and decode speech as shown in Figure 11. The approach assumes that frequency, bandwidth and intensity of spectral mixtures can be used to encode speech. The parameters can be used to reconstruct the synthesized version for approximating the spectrum.

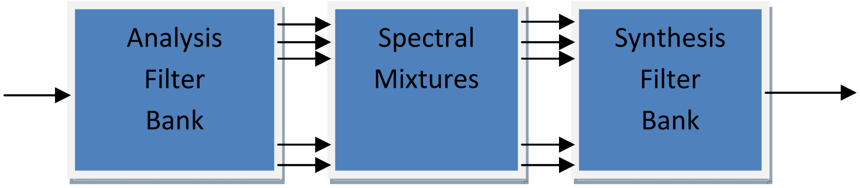

Figure 11 Diagram of the Proposed Vocoder Implementation

Instead of estimating the vocal track and the excitation, the signal is represented as a set of three parameter vectors: the center frequency, bandwidth and power of the spectral response. Synthesized speech using Gaussian distribution results in Gaussian pulses in a time with a width that is inversely proportional to the bandwidth. As an alternative to Gaussian pulses, we can synthesize an approximate the time response with sinc functions shown in equation 3.2.

$$p(f|z;\vec{\theta}) \approx \left|\frac{sinc(\pi b(f - f_c))}{\pi f}\right| \quad 3.2$$

The filter banks were implemented in Matlab®. Figure 12 shows the decomposition and synthesis of a 256-channel filterbank. The filter uses a Kaiser window with 30 dB of attenuation and transition bandwidth of 5 percent. These parameters were chosen to preserve the reconstructed signal from distortion.

The advantage of this approach is the little complexity in the implementation. The scheme does not need voice/unvoiced detection.

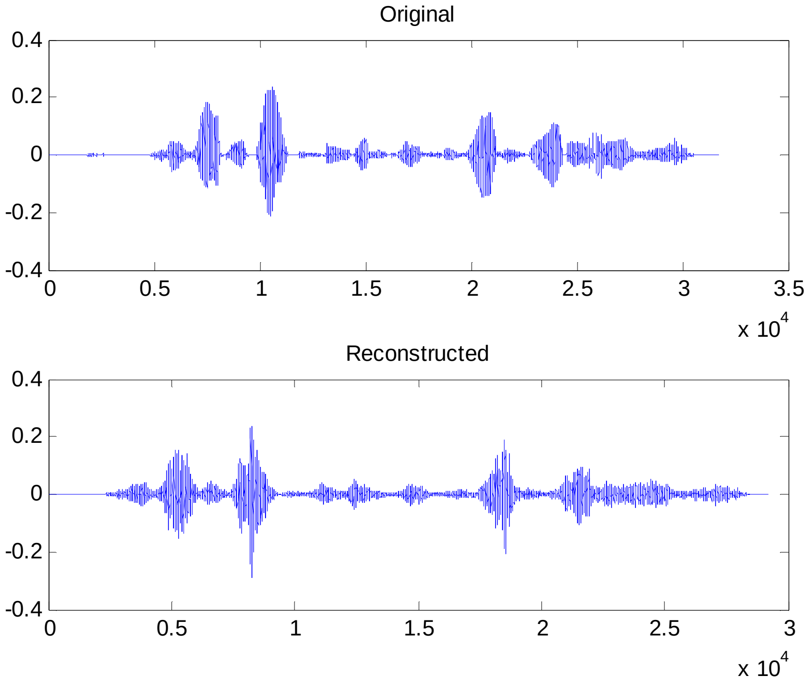

Figure 12 Original and Decoded Speech

Figure 13 shows the approximation of the voice and unvoiced speech. The algorithm provides a crude approximation in both cases.

The algorithm is implemented using matrices. The computation of the posterior distribution in the E-Step requires $O(N * M * I)$ floating point operations, where N is the number of data points, M is the number of mixtures and I is the number of iterations used. In this particular example the number of floating points operations is approximately $O(N * M * I) = 256 * 10 * 40 = 102400$.

The achieved compression is as follows: The speech audio file is 57KB sampled at 8kHz. The speech separated into its frequency components using a 256-channel filter bank. The spectral mixture approach uses 15 Gaussian Mixtures and produces three outputs: center frequency, bandwidth and power composition. The data is saved using the following fields:

1. 1 byte: center frequency, the data range is (0-255)
2. 1 byte: bandwidth, the data range is (0-255)
3. 1 byte: power composition, the data range is (0-100)
4. 8 bytes: total power per window of data, the data range can currently take floating point numbers of double precision. This value is necessary because the spectral mixture algorithm only takes normalized distributions. The value could be reduced to 4 bytes.

The 57 KB file produces is compressed to a 4.27 KB file. The compression ratio is 12.12. The file plays for 3 seconds. The original bit rate is 57K * 8 /3 = 152 kbps. The compressed audio rate is 4.27K * 8/3 = 11 kbps.

For having a simple architecture the proposed vocoder appears to have a fair performance with some hoarse and buzzing sound. Of course, there are state of the art vocoders that work much better at lower rates, for example, the MBE vocoder works at 4.15 kbps and the MELP works at 2.4 kbps. [10] Nevertheless, these vocoders have sophisticated processing and it is not the intention of this project to beat the current state of the art.

3.3.Other Applications

3.3.1. Blind Deconvolution

One of the original goals was to develop a deconvolution model that would provide estimates of the periodicity of a cyclostationary process.

The model used was a distribution shown in equation 3.3. The equation describes a square pulse modulated by a cyclostationary signal with an offset. The cyclostationary signal is created with a raised cosine filter.

$$e^{-1.7^N (\frac{y}{L})^N} \left\{ \begin{array}{ll} 1+\alpha \dfrac{\cos\left(\frac{\pi y \beta}{T}\right) \operatorname{sinc}\left(\frac{\pi y}{T}\right)}{1-\frac{4y^2\beta^2}{T^2}} & y \neq \dfrac{T}{2\beta} \\ 1+\alpha \dfrac{1}{2} \beta \sin\left(\dfrac{\pi}{2\beta}\right) & y = \dfrac{T}{2\beta} \end{array} \right\} \qquad 3.3$$

The idea was to model a cyclostationary signal as distribution with parameters β, L and T. First, a raised cosine function is created. The response adds an offset. The signal is modulated with a square pulse. (Figure 13) The separation between the impulses was chosen to be the same as the distance between the peak of the filter response and the null. This condition is required for interpolating data points between impulses.

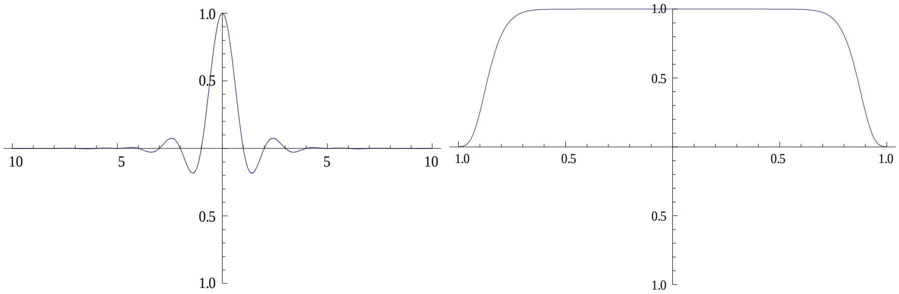

Figure 13 Cyclostationary signal and square pulse

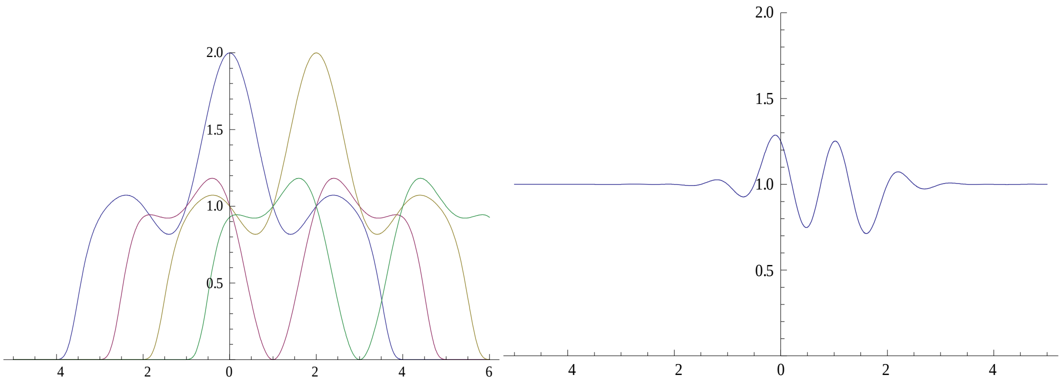

Figure 14 Mixtures for deconvolution and signal with offset

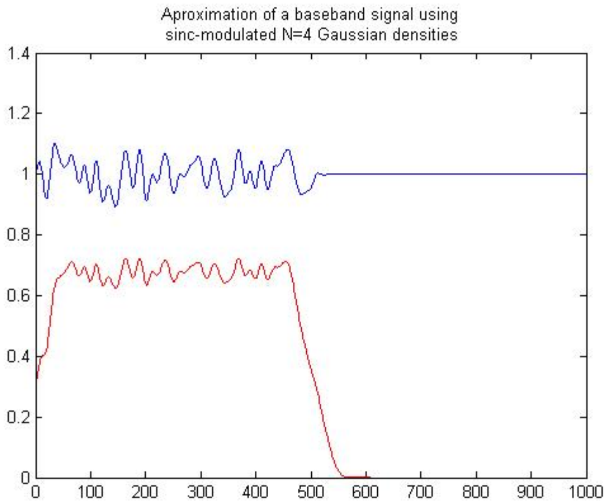

Figure 15 Synthesis of a cyclostationary signal with offset

The cyclostationary signal is transformed into a mix of density functions. (Figure 15) The approach has problems that were found to be impossible to overcome. The mixing of mixtures creates a signal that is extremely difficult to separate. For instance, the proximity of two distributions can cause the algorithm to converge to a local maximum instead of the absolute maximum. This is a weakness of all the algorithms based on Expectation Maximization. Second and probably the most critical consideration is given by equation 2-57. The algorithm is an iterative process that reduces the divergence between the red trace and the blue trace in Figure 15; however, there is no guarantee that the minimum divergence produces the optimal parameters for this distribution. The idea was abandoned due to these two difficulties.

3.3.2. Image Detection

An image can be represented as a mixture of densities. So one natural question would be whether the Spectral Mixture approach can use as an image detector.

The development is not so different from the development of the affine transform model shown in equations 2.19 through 2.20. The development produces factors that depend on the gradients. This is interesting because gradients are used in image processing for shape detection. Other than a mathematical curiosity, the scheme fails due to the presence of multiple local maxima and the number of degrees of freedom, i.e., the dimension of the parameter vector. This idea was abandoned due to these two difficulties.

3.3.3. ABC Process

The Adjustable Bandwidth Concept (ABC) is a methodology developed by [7] that can be applied for the detection of wideband signals. The method process creates averages and high frequency versions (wavelet decomposition) in the frequency spectrum. This research explored the possibility of combining both approaches.

The ABC algorithm performs much better with another algorithm called Connective Components. The latter algorithm works very efficiently with the binary data produced by the ABC method. We discarded using the Spectral Mixture for this application due to the computational efficiency of the Connective Component.

4. Conclusions

In report AFRL-RI-RS-TR-2008-266, it was noticed that raising the likelihood to an integer power $S(f)$ was equivalent to adding $S(f)$ samples with the same value. This is true in both: the maximum likelihood approach and the Expectation Maximization method. This fact was use to develop an inference method for histograms. The method was referred as Spectral Mixtures.

In this report, the theory of Maximum Likelihood and Expectation Maximization methods was reviewed. For developing the Expectation Maximization method, we took advantage of the posterior density from equation 2.5 to produce an average over a modified likelihood given by equation 2.7. The maximization process resulted in the Expectation Maximization algorithm.

The development of the Spectral Mixture Model considered the quantization of the sample space in histogram bins as shown in equation 2.11. There is a variable associated with this quantization: this is the histogram count $S(f)$. This quantity was constrained to the set of positive integers, but this constraint can be relaxed to any positive real number. The quantity $S(f)$ can be expressed as a probability distribution, which makes possible to reformulate the Spectral Mixture Model approach as a minimization of the divergence of two distributions. The new algorithm takes the form of equation 2.58 where $\vec{\theta}$ is the parameter vector

$$\vec{\theta}_{new} = arg \min_{\vec{\theta}} D_{KL}\left(q(\vec{x},z;\vec{\theta}_{old})\,||\,p(\vec{x},z;\vec{\theta})\right) \quad 4.1$$

and $q(\vec{x},z;\vec{\theta}_{old})$ and $p(\vec{x},z;\vec{\theta})$ are given by:

$$q(\vec{x},z;\vec{\theta}_{old}) = \sum p(\vec{x}|\bar{z};\vec{\theta}_{old})\bar{p}(\vec{x})$$

$$p(\vec{x},z;\vec{\theta}) = \sum p(\vec{x}|\bar{z};\vec{\theta})p(\bar{z}).$$

$$\bar{p}(\vec{x}_m) = S(\vec{x}_m)/\int S(\vec{x})dV$$

The Spectral Mixture algorithm was applied to the problem of spectral sensing of communication signals and speech signals. In both cases, the amplitude of the spectrum is considered while the angle is neglected. The method works fine for cases where no or little noise floor is present. In order to overcome this difficulty, one of the mixtures is forced to have constant variance. The variance is made large enough, so the mixture converges to the noise floor as shown in Figure 8. Clipped Gaussian densities were used for simplification purposes. The log-likelihood of such distributions is reduced to a polynomial expression and the maximization process becomes less complex. Sometimes, we get closed form solutions, but this is not always the case.

The Spectral Mixture was applied to deconvolution and image processing. It was found that the Spectral Mixture approach inherits all the weaknesses of the Expectation Maximization. One weakness is the resolution of two mixtures with a small separation. The algorithm tends interpret both mixtures as one mixture. The other problem is the convergence to a local maximum. Sometimes, the local maximum solution does not represent the best approximation. The third problem is the degrees of freedom. Adding more degrees of freedom, i.e., adding more parameters makes the problem more difficult to solve due to the increase of complexity.

As a final conclusion, equation 4.1 is a generalization of the Expectation Maximization algorithm. Under special cases, the algorithm is reduced to K-Mean and Parzen window.

5. References

[1] S. Bernadin, Wavelet Processing for Pitch Estimation, Proceedings of the 38th Southeastern Symposium on System Theory. IEEE, 2006

[2] C. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[3] W. Chu, Speech Coding Algorithms, Wiley & Sons, 2003

[4] J. Deller et Al., Discrete-Time Processing of Speech Signals, IEEE Press, 2000T.

[5] R. Duda, Pattern Classification, 2nd Edition, John Wiley and Sons, 2001

[6] T. Dutoit, M. Terran, Applied Signal Processing, Springer, New Jersey 2009

[7] Galleani, L. Cohen, A. Noga, A Time-Frequency Approach to the Adjustable Bandwidth Concept, Elsevier, Digital Signal Processing, Vol. 16, Issue 5, September 2006

[8] D. Griffin, J. Lim, Multiband Excitation Vocoder, IEEE Transactions on Acoustic, Speech and Signal Processing, Vol 36, 1988

[9] A. Hero, et Al, Alpha-Divergence for Classification, Indexing and Retrieval, Communication and Signal processing Laboratory, Technical Report CSPL-328, May 2001.

[10] Z. Li, M. Drew, Fundamentals of Multimedia, Prentice Hall, NJ 2004

[11] M. Milosavljevic et Al., Robust LPC Parameter Estimation in Standard CELP 4800 b/s Speech Coder, IEEE TENCON – Digital Signal Processing, 1996

[11] T. Moon, The Expectation Maximization Algorithm, IEEE Signal Processing Magazine; Vol 13, No. 6, November 1996, pg. 47-64.

[13] E. Parzen, On Estimation of a Probability Density Function and Mode, Ann. Math. Stat. 33, pp. 1065.

[14] S. Theodoridis, K. Koutroumbas; Pattern Recognition 3rd Edition, Elsevier, 2006

[15] A. Vega-Irizarry, Automated Spectral Survey Techniques for Blind Demodulation/ Modulation Classification, AFRL-RI-RS-TR-2008-266, October 2008

[16] S. Vaseghi, Multimedia Signal Processing